

CAPÍTULO 2

EL VALOR AGREGADO COMO UNA METODOLOGÍA DE ANÁLISIS DE DESEMPEÑO FORMATIVO

2.1. ANÁLISIS DE LA METODOLOGÍA DE VALOR AGREGADO

A continuación se analizan los estudios que han empleado el análisis de bases de datos como insumo para la metodología asociada al valor agregado, y cuáles han sido los usos que se le han dado a esta información en el país (tabla 2).

Tabla 2. Análisis de investigaciones en temas educativos analizadas y limitaciones que se han presentado

Autores	Datos utilizados	Principales conclusiones	Restricciones de la muestra	Otros comentarios
Gaviria y Barrientos (2001)	Pruebas Icfes de Bogotá de 1999	1) La educación de los padres tiene un efecto sustancial en el rendimiento académico, 2) efecto que se transmite principalmente a través de la calidad de los planteles educativos, 3) los cuales inciden de manera notable sobre el rendimiento.	Solo se tuvieron en cuenta los individuos matriculados en grado 11 durante 1999. Excluidos aquellos que presentaron la prueba algunos años después de terminar la secundaria como quienes lo hicieron después de validar el bachillerato, individuos mayores de 30 años y quienes obtuvieron un puntaje nulo en la totalidad del examen o en los componentes de lenguaje y matemáticas.	Pruebas Icfes fueron modificadas en el año 2000.

4 Actualmente es el director general de la Corporación Penser. Ha sido responsable de procesos de aseguramiento de la calidad y de acreditación internacional, asesor de la Rectoría y del Consejo Superior y vicerrector académico. PhD en Medición y Evaluación en Educación por la Universidad de Montreal. Es par evaluador de Minciencias y de programas e instituciones del CNA. Además, es *senior member* del Institute of Industrial & Systems Engineers. Ha publicado diversos trabajos y artículos sobre medición y evaluación de programas e instituciones. Campos de trabajo: evaluación educativa, aseguramiento de la calidad y gestión del conocimiento.

5 Economista con especializaciones en Economía y Gerencia Social. Maestría en Economía y estudios de Doctorado en la Universidad de Sevilla (España). Actualmente es docente y coordinador de posgrados de la Facultad de Ciencias Administrativas en la Universidad Autónoma de Occidente. Alterna el ejercicio académico con la consultoría empresarial en temas de calidad.

Autores	Datos utilizados	Principales conclusiones	Restricciones de la muestra	Otros comentarios
Gaviria y Barrientos (2001)	Pruebas Icfes de Bogotá de 1999, características de los planteles educativos Bogotá de 1999 (DANE)	1) Es posible aumentar el rendimiento académico por medio de inversiones en la capacitación de los docentes y en la infraestructura del plantel, 2) pero el impacto de estas inversiones está sujeto a modificaciones en los incentivos y las prácticas pedagógicas concomitantes.	Se encontró información detallada sobre las características del plantel para 693 de los 1105 planteles de la base de datos del Icfes.	Se hizo una comparación entre la base de datos del Icfes y la base de datos formada con los 693 planteles, para comprobar que la exclusión de algunas instituciones no generó sesgos de selección.
Barón (2010)	Pruebas realizadas por el Icfes (Saber 5, 9, 11 y Ecaes) en 2009	Los estudiantes de la ciudad de Barranquilla presentan un rendimiento académico inferior al de los estudiantes de ciudades como Bogotá y Bucaramanga. Estas brechas de rendimiento son importantes y persistentes a través del proceso de formación de capital humano (grado quinto, noveno, undécimo y superior).	Se cuantifica las brechas de rendimiento y las causas de estas diferencias entre Barranquilla y Bogotá para el caso específico de la prueba Saber-11 (Icfes). No hubo restricciones en la muestra pues se trabajó con la base de datos del Icfes.	Usando una descomposición contrafactual, el análisis se enfoca en qué tanto de la brecha de rendimiento en la ciudad se puede explicar por las diferencias en las características de las instituciones y las diferencias en las características de los estudiantes y sus familias.

Autores	Datos utilizados	Principales conclusiones	Restricciones de la muestra	Otros comentarios
OCDE (2011)	Se presentan ejemplos de pruebas y resultados obtenidos en países que hacen parte de la organización.	Al final del Capítulo 3 se muestra el “Establecimiento de un sistema basado en modelos de valor agregado: pasos básicos en la fase de implementación”.	La medición del aprendizaje de los alumnos. Mejores prácticas para evaluar el valor agregado de las escuelas.	Se centra en aspectos de su diseño, el tipo de modelos posibles y los aspectos estadísticos y metodológicos pertinentes. La parte III se centra en la puesta en marcha de sistemas basados en modelos de valor agregado, y abarca tanto los aspectos políticos e institucionales que deben abordarse como las consideraciones técnicas que necesitan resolverse.
Monroy-Mateus et al. (2018)	Resultados de las pruebas Saber 11 y Saber PRO y contraste a través de grupo de referencia.	Se conoce el aporte que genera una IES, sus principales competencias y las que necesitan mejora. Se detectan problemas institucionales colectivos y por programas. Se correlacionan variables asociadas al desempeño. Se sugieren estrategias y planes de mejoramiento.	Que el Icfes facilite los datos a analizar y correlacionar y que estos se puedan cotejar en las dos pruebas empleadas.	Se identifica el valor agregado a partir del análisis del desarrollo de las competencias de los estudiantes, con la intención de crear estrategias en pro del mejoramiento continuo.

Fuente: elaboración propia con base en apuntes de clase de Barón (2010), Organización para la Cooperación y el Desarrollo Económicos (OCDE, 2011) y Monroy-Mateus et al. (2018).

Según estudios de la OCDE (2011), Bogoya et al. (2017), Koc (2018), Cassandra (2018) y Franco-Gallego (2019), los modelos de valor agregado son una mejora sustancial en comparación con muchas medidas de desempeño escolar que actualmente se utilizan.

Las comparaciones de puntuaciones brutas de pruebas proporcionan cierta información importante, pero son medidas deficientes del desempeño escolar.

Por su parte, el Instituto Colombiano para la Evaluación de la Educación (Icfes, 2019) establece que los modelos de valor agregado son modelos estadísticos que buscan cuantificar cuánto aportan las instituciones educativas a sus estudiantes en términos de aprendizaje. Esto teniendo en cuenta factores institucionales e individuales, por lo que actualmente estos modelos son una herramienta convencional en la literatura internacional sobre medición de la calidad de la educación.

Así es como el análisis de valor agregado, según los anteriores autores, toma en cuenta el rendimiento previo y se generan resultados que en gran medida reflejan diferencias en las características contextuales, como los antecedentes socioeconómicos de los alumnos. Los modelos de aprovechamiento contextualizado intentan abordar estos problemas al medir el impacto de las características contextuales sobre una medida de desempeño específica, pero son menos útiles para separar los efectos escolares en el avance estudiantil de otras características contextuales y, por ende, son menos útiles para medir el desempeño escolar.

Los modelos de valor agregado pretenden resolver estos problemas, al incorporar medidas de aprovechamiento previo y, en algunos casos, características contextuales de los alumnos (Thieme et al., 2011; Isázaiga-David y Gabalán-Coello, 2014). Esto permite un análisis más refinado del progreso del desempeño de los alumnos y resulta más eficaz para desentrañar los efectos de diversos factores que afectan al avance de los alumnos. Estas ventajas permiten una mayor precisión al medir el desempeño, lo que genera más confianza en la interpretación de las mediciones de desempeño escolar.

Para aquellos países y sistemas educativos que se encuentran en procesos de reformas e innovación, se sostiene que los modelos de valor agregado pueden contribuir a los siguientes esfuerzos, según los autores señalados:

- Aprendizaje en todos los niveles del sistema educativo, al identificar con precisión los aspectos de mayor y menor desempeño.
- Identificación y análisis de “lo que funciona” para procesos de mejora escolar.
- Establecimiento de sistemas de rendición de cuentas más transparentes y más equitativos que pueden servir como base para el diseño de incentivos para elevar el desempeño escolar.

- Desarrollo de sistemas de información que permitan a las escuelas analizar y evaluar su desempeño y fortalecer el sistema de evaluación.
- Mayor eficacia y equidad en la asignación de recursos (humanos, materiales y económicos) a las áreas críticas que presenten mayor necesidad.
- Atención a desigualdades socioeconómicas arraigadas que puedan estar ocultas en el ámbito escolar, dadas por medidas de desempeño indiscriminadas, imprecisas e inequitativas.

Una característica distintiva de los modelos de valor agregado es la inclusión de medidas anteriores de desempeño, que permiten una estimación más precisa de la contribución de la escuela al progreso de los alumnos. Doran e Izumi (2004), citados en OCDE (2011), destacaron las ventajas de los modelos de valor agregado al registrar a los alumnos a través del tiempo en comparación con modelos transversales (o de aprovechamiento contextualizado), que ofrecen una imagen “estática” del desempeño escolar.

Los modelos de valor agregado registran de mejor manera, según la OCDE (2011), aquellos factores no observados que influyen en la medición de desempeño inicial (la capacidad del alumno, por ejemplo), lo cual representa un problema sistémico en la mayoría de los modelos de aprovechamiento contextualizado. Posición compartida por Soares et al. (2017). Así, se propone un indicador para medir el valor agregado por las regiones de los consejos regionales de planificación económica y social (Corpes), llamado *porcentaje de efectividad máxima*, el cual se presenta a continuación.

2.2. PORCENTAJE DE EFECTIVIDAD MÁXIMA

La primera parte de este capítulo pretende mostrar los desempeños estudiantiles en contexto. Esto resalta la importancia de poder llevar todo a un mismo referente que coadyuve a la identificación de elementos de comparabilidad, pero lógicamente teniendo en cuenta las variaciones y las distribuciones asociadas a cada variable objeto de análisis. En este sentido, es una propuesta metodológica que, como toda propuesta académica, se espera que se encuentre en constante discusión y ajuste para posteriores investigaciones en la temática.

Esto significa que no bastaría con los rendimientos absolutos, dado que las condiciones de la población marcan una tendencia importante, y sobre la cual se deben analizar los desempeños individuales. Por lo anterior, se usarán los porcentajes de efectividad máxima como muestra relativa de los puntajes. El término *porcentaje de efectividad máxima* es una construcción terminológica (propuesta), responsabilidad exclusiva de los autores,

en la cual se intentan describir los puntajes obtenidos por los estudiantes en función del máximo alcanzable para cada cohorte de evaluación (véanse estudios anteriores de Isáziga-David y Gabalán-Coello, 2014; Isáziga-David et al., 2014).

Desde esta perspectiva, es una razón que tiene en cuenta en el numerador el puntaje promedio alcanzado por cada estudiante en la prueba respectiva y en el denominador el puntaje máximo que podría obtener cualquier estudiante en la prueba teniendo en consideración los promedios nacionales para la fecha en la cual el estudiante presentó la prueba (análisis por cohortes) más tres veces la desviación estándar de la prueba en la misma fecha (análisis por cohortes).

Dicha razón se ampara en la regla empírica de la distribución normal (Lind et al., 2012), en la cual la gran mayoría de los desempeños se encontrarán entre el promedio y más o menos tres desviaciones estándar; por tanto, una aproximación a los máximos valores estaría concentrada en analizar el promedio más las tres desviaciones estándar de cada cohorte o periodo en la cual se presentaron las pruebas censales.

Por este motivo, se toman las regiones Corpes que se van a analizar y se emplean los puntajes obtenidos por los estudiantes en el componente “Comprensión lectora”. Estos puntajes se llevan a una expresión relativa, como lo es el porcentaje de efectividad máxima. En este caso, se tendrían los desempeños en el contexto del total de estudiantes que presentaron dicha prueba en el país.

La otra medición de este mismo individuo (muestra pareada) se obtiene a partir de una visión retrospectiva, en la cual se busca para cada estudiante su desempeño en el componente “Lenguaje” cuando presentó el Saber 11 — véase el sustento metodológico en Gabalán-Coello y Vásquez-Rizo (2016)—. Igual que en el procedimiento anterior, se expresan estos rendimientos de una manera relativa a través del concepto de porcentaje de efectividad máxima (figura 1).

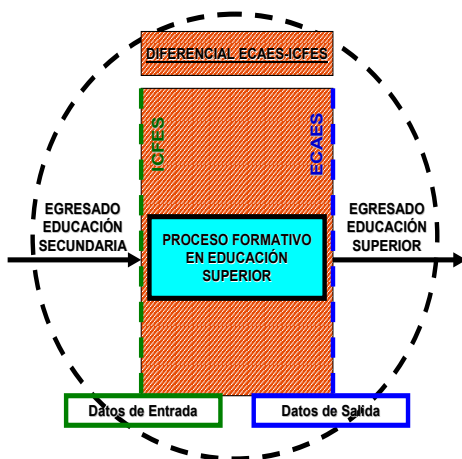


Figura 1. Fuentes e insumos del análisis

Fuente: elaboración propia.

Se muestran las diferencias entre los porcentajes de efectividad máxima antes del proceso formativo y después de este, esperando que exista un mejoramiento en el posicionamiento relativo de los estudiantes luego de tener una formación universitaria. En los resultados se describe la proporción de estudiantes que igualan o superan su porcentaje de efectividad máxima (superan sus desempeños relativos), de la misma manera que su complemento evidenciará aquellos que disminuyeron sus posiciones relativas. La proporción de estudiantes que igualan o superan marcará un indicador positivo en la medida en que se aproxime al 100 %, situación que permitirá empíricamente mostrar una incidencia del proceso formativo en los procesos de calidad del individuo.

La proporción de estudiantes que igualan o superan su porcentaje de efectividad máxima (Saber PRO-Saber 11) se define de acuerdo con la siguiente notación:

$$P = \frac{\sum pemsaberpro \geq pemsaber1}{n}$$

Donde:

pemsaberpro = porcentaje de efectividad máxima en Saber PRO.

pemsaber1 = porcentaje de efectividad máxima en Saber 11.

n = número total de estudiantes de la región Corpes analizada que presentaron Saber PRO.

La sumatoria permite contabilizar el número de estudiantes en los cuales los porcentajes de efectividad máxima en la prueba Saber PRO son mejores o iguales que los porcentajes de efectividad máxima de la prueba Saber 11.

Por otra parte, también se emplea un indicador utilizado con frecuencia: “promedio del porcentaje de efectividad máxima en Saber PRO DE LA COHORTE DE SALIDA / PROMEDIO DEL PORCENTAJE DE EFECTIVIDAD MÁXIMA EN Saber 11 de cohorte de salida al momento de ingreso al programa”. En esta medición, una relación mayor que uno indicará que el programa ayudó a elevar la posición relativa del grupo frente al promedio nacional y, por tanto, agregó valor. Una relación por debajo de la unidad, indica lo opuesto.

La notación se establece de la siguiente manera, con respecto a este indicador:

$$I = \frac{\frac{\sum_{i=1}^n pemsaberpro_i}{n}}{\frac{\sum_{i=1}^n pemsaber1_i}{n}}$$

En la metodología propuesta se continúan explorando evidencias que lleven a pensar en el progreso relativo de los estudiantes y la incidencia del proceso formativo en sus posiciones relativas. Por tanto, se plantean tres comparaciones estadísticas donde se desea conocer si existen progresos significativos, acorde con la distribución de los puntajes, entre las puntuaciones antes de la intervención educativa y después de esta. Para ello, se emplea la prueba para muestras pareadas: la prueba t (supuesto distribución normal), la prueba de Wilcoxon y la prueba de signos (sin supuesto sobre la distribución).

2.3. PRUEBA DEPENDIENTES O PAREADAS

En algunas oportunidades se tienen muestras dependientes o “pareadas”, ya sea porque se realizaron dos observaciones sobre el mismo elemento o individuo o porque se hizo un pareo de elementos de acuerdo con ciertas características (Marín-García et al., 2014). En estos casos lo que se tiene, realmente, es una muestra de “pares” de elementos, y lo que interesa es determinar si, en promedio, las diferencias entre esos pares difieren significativamente o no de cero.

2.3.1. PRUEBA T PARA MUESTRAS RELACIONADAS

Tomando como referencia el *software* SPSS, a partir de la *Guía para el análisis de datos* (Universidad de Cádiz, 2019), se tiene que la prueba t para dos muestras relacionadas permite contrastar hipótesis referidas a la diferencia entre dos medias relacionadas.

Ahora se dispone de una población de diferencias con media μ_D , obtenida al restar las puntuaciones de mismo grupo de casos en dos variables diferentes o en la misma variable medida en dos momentos diferentes (de ahí que hablemos de muestras relacionadas). De esa población de diferencias se extrae una muestra aleatoria de tamaño n y se emplea la media \bar{Y}_D de esas n diferencias para contrastar la hipótesis de que la media μ_D de la población de diferencias vale cero.

Desde el punto de vista estadístico, este contraste es idéntico al de prueba t para una muestra. La única diferencia existente entre ambos contrastes es que ahí se tenía una muestra de puntuaciones obtenida al medir una sola variable y ahora se cuenta con dos muestras relacionadas (o una muestra de pares de puntuaciones) que se convierte en una sola muestra de diferencias restando las puntuaciones de cada par.

El estadístico o prueba t sigue siendo una tipificación de la media muestral de las diferencias \bar{Y}_D (S_D se refiere a la desviación típica insesgada de las n diferencias):

$$T = \frac{\bar{Y}_D - \mu_D}{\hat{\sigma}_{\bar{Y}_D}} = \frac{\bar{Y}_D - \mu_D}{S_D / \sqrt{n}}$$

Este estadístico t se distribuye según el modelo t de Student con $n-1$ grados de libertad y, por tanto, permite conocer la probabilidad asociada a los diferentes valores \bar{Y}_D que es posible obtener en muestras aleatorias de tamaño n .

Al igual que antes, para que el valor t se ajuste apropiadamente al modelo de distribución de probabilidad t de Student es necesario que la población de diferencias sea normal. No obstante, con tamaños muestrales grandes, el ajuste del estadístico t a la distribución t de Student es lo suficientemente bueno incluso con poblaciones originales alejadas de la normalidad.

2.3.2. PRUEBA DE WILCOXON

Tomando como referencia *SPSS: Guía para el análisis de datos* (Universidad de Cádiz, 2019), se tiene que esta es una prueba utilizada en el análisis no paramétrico que permite analizar datos de diseños con medidas repetidas.

Se supone tomar dos medidas (X_i e Y_i) a un grupo de m sujetos y que calculamos las diferencias en valor absoluto entre las dos puntuaciones de cada par:

$$D_i = |X_i - Y_i| \quad (i = 1, 2, \dots, m)$$

Se descartan las D_i nulas y se considerarán únicamente las n diferencias D_i no nulas ($n \leq m$). Asignemos rangos (R_i) desde 1 hasta n a esas D_i no nulas: el rango 1 a la D_i más pequeña, el rango 2 a la D_i más pequeña de las restantes..., el rango n a la D_i más grande (si existen empates, se resuelven asignando el promedio de los rangos). Se suma ahora, por un lado, los R_i^+ , es decir, los rangos correspondientes a las D_i con $X_i > Y_i$, y llamar S_+ a esta suma; y sumar, por otro lado, los R_i^- , es decir, los rangos correspondientes a las D_i con $X_i < Y_i$, y llamemos S_- a esta otra suma. Si se supone que las puntuaciones X_i e Y_i proceden de poblaciones con la misma mediana ($Mdn_x = Mdn_y$), se debe esperar que:

$$P(X_i < Y_i) = P(X_i > Y_i)$$

Por lo que si la hipótesis $H_0: Mdn_x = Mdn_y$ es verdadera, en una muestra aleatoria de n observaciones se deben encontrar aproximadamente tantos valores $X_i > Y_i$ como valores $X_i < Y_i$ (salvando, por supuesto, las fluctuaciones atribuibles al azar muestral). Además, si la distribución de las diferencias es simétrica (lo cual exige escala de intervalo o razón), las D_i positivas se alejarán de cero en igual medida que las D_i negativas, de donde es fácil deducir que:

$$S_+ = \sum R_i^+ \approx S_- = \sum R_i^-$$

Es decir, si $Mdn_x = Mdn_y$, la distribución de las diferencias D_i es simétrica, S_+ y S_- tomarán valores parecidos. Por tanto, una fuerte discrepancia entre S_+ y S_- hará dudar

de la veracidad de H_0 . De modo que se procede utilizar los valores S_+ y S_- para obtener información sobre la hipótesis $H_0: Mdn_x = Mdn_y$ (Wilcoxon, 1945; Litchfield y Wilcoxon, 1949).

Con tamaños muestrales pequeños no resulta complicado obtener la distribución exacta del estadístico S_+ (o S_-). Pero es más rápido obtener una tipificación de S (S se refiere al menor de S_+ y S_-) cuya distribución se aproxima, conforme el tamaño muestral va aumentando, al modelo de probabilidad normal $N(0,1)$ (k se refiere al número de rangos distintos en los que existen empates y t_i al número de puntuaciones empatadas en el rango i).

$$Z = \frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^k \frac{t_i^3 - t_i}{48}}}$$

El SPSS ofrece el nivel crítico bilateral resultante de multiplicar por dos la probabilidad de obtener valores menores o iguales que Z .

2.3.3. PRUEBA DE LOS SIGNOS

Tomando como referencia SPSS: *Guía para el análisis de datos* (Universidad de Cádiz, 2019) se describe esta prueba, como una utilizada en el análisis no paramétrico que permite analizar datos de diseños con medidas repetidas. Al igual que la prueba de Wilcoxon, la prueba de los signos permite contrastar la hipótesis de igualdad entre dos medianas poblacionales. Pero mientras la prueba de Wilcoxon aprovecha la información ordinal de los datos (aunque exige nivel de medida de intervalo o razón), la prueba de los signos solo aprovecha los datos de sus propiedades nominales (aunque exige nivel de medida al menos ordinal). Se supone que tomando dos medidas (X_i e Y_i) a un grupo de m sujetos, se calculan las diferencias en valor absoluto entre las dos puntuaciones de cada par:

$$D_i = |X_i - Y_i| \quad (i = 1, 2, \dots, m)$$

Desechemos las D_i nulas y consideremos únicamente las n diferencias D_i no nulas ($n \leq m$). Si supone que las puntuaciones X_i e Y_i proceden de poblaciones con la misma mediana ($Mdn_x = Mdn_y$), debe verificarse que:

$$P(X_i < Y_i) = P(X_i > Y_i) = 0,5$$

De modo que si la hipótesis $H_0: Mdn_x = Mdn_y$ es verdadera, en una muestra aleatoria de n observaciones se deben encontrar aproximadamente tantos valores $X_i > Y_i$ como valores $X_i < Y_i$, es decir, aproximadamente tantas diferencias D_i positivas como negativas (salvando, por supuesto, las fluctuaciones atribuibles al azar propio del proceso de muestreo). Bajo estas condiciones, las variables:

n_+ = número de signos positivos

n_- = número de signos negativos

Se distribuyen según el modelo binomial con parámetros n y $\pi = 0,50$. De modo que se puede utilizar la distribución binomial para conocer las probabilidades asociadas a n_+ y n_- y, basándonos en ellas, contrastar la hipótesis $H_0: Mdn_x = Mdn_y$.

Si $n \leq 3$, el SPSS toma el valor $k = \min(n_+, n_-)$, y utilizando las probabilidades de la distribución binomial, calcula el nivel crítico bilateral resultante de multiplicar por dos la probabilidad de obtener valores iguales o menores que k .

Si $n > 24$, el SPSS tipifica el valor de k (utilizando corrección por continuidad) y ofrece el nivel crítico resultante de multiplicar por dos la probabilidad de encontrar valores iguales o menores que Z :

$$Z = \frac{k + 0,5 - n/2}{0,5\sqrt{n}}$$

2.4. FUENTES DE INFORMACIÓN

Las fuentes de información que se emplean en esta parte del libro serán las bases de datos de las pruebas Saber 11 y Saber PRO, aplicadas en el periodo comprendido entre 2000 y 2009. Se toma una muestra y se aparean los datos de los estudiantes por las variables documento de identidad, nombre, institución educativa, departamento y región Corpes. Se debe tener presente que los datos para el actual trabajo deben permitir la trazabilidad del estudiante en el periodo de estudio.

En esta dirección, y a efectos de una mayor claridad en lo expuesto en el presente libro y en la ventana de observación y análisis tomada, se debe decir que a partir de 2000 las pruebas censales colombianas se orientaron hacia la evaluación por competencias, en lugar de la evaluación de conocimientos y aptitudes. Esto se hizo en concordancia con lo establecido por los lineamientos curriculares y, posteriormente, por los estándares básicos de competencias del Ministerio de Educación Nacional (MEN, 2006), con el objetivo de consolidar un Sistema Nacional de Evaluación Estandarizada que consiguiera la alineación de todos los exámenes que lo conforman (al respecto, véase Icfes, 2013).

De esta manera, se modificó la estructura del examen Saber 11 y, posteriormente, de nuevo a partir del segundo semestre de 2014, para que sus resultados fueran comparables, en términos de la evaluación de competencias genéricas, con los de otras pruebas del Sistema Nacional de Evaluación Estandarizada, como son las pruebas Saber 3, 5 y 9, y el examen de educación superior, Saber PRO. Esta alineación implicó que los exámenes debían estar articulados en torno a la evaluación de unas mismas competencias en diferentes grados de desarrollo, lo que permitió pasar de un sistema con mediciones aisladas en la educación media a uno que hiciera un seguimiento sistemático de los resultados de la educación a través de diferentes niveles (Icfes, 2013).

La tabla 3 muestra cómo esta alineación reestructuró el examen en torno a la evaluación de competencias genéricas. En primer lugar, con la creación de la subprueba de “Competencias ciudadanas”; en segundo lugar, diferenciando en la prueba de “Matemáticas”, una subprueba de “Razonamiento cuantitativo” y, finalmente, fusionando pruebas en torno a competencias genéricas que evalúan elementos comunes (los componentes de “Lenguaje” y “Filosofía” se fusionaron en la prueba de “Lectura crítica”; mientras que los componentes de “Física”, “Química” y “Biología” se fusionaron en la prueba de “Ciencias naturales”, y las “Competencias Ciudadanas” se incorporaron a través de una prueba de “Sociales y ciudadanas”) (Icfes, 2013).

Tabla 3. Histórico de la alineación de los componentes de las pruebas en torno a la evaluación de competencias genéricas

2000-I a 2005-2	2006-I a 2014-I	2014-II en adelante
Lenguaje Filosofía	Lenguaje Filosofía	Lectura crítica
Matemáticas	Matemáticas	Matemáticas (Razonamiento cuantitativo)
Física Química Biología	Física Química Biología	Ciencias naturales
Historia	Ciencias sociales (Historia-Geografía)	Sociales y ciudadanas
Geografía		

Fuente: Icfes (2013).

Ya en 2003, el Icfes inició la evaluación del ciclo de educación superior mediante la aplicación de exámenes específicos dirigidos a diversos programas académicos, llamado Examen de Estado de Calidad de la Educación Superior (Ecaes), con la evaluación de 22 programas, hasta llegar a 55 programas en 2007. En ese periodo, el examen se centró en competencias específicas por programa y no todos los estudiantes presentaban la prueba, debido a que no tenía un carácter obligatorio, pues no se contaba con exámenes para cada uno de los programas existentes, situaciones que dificultaron hacer un análisis de componentes transversales.

Fue entonces para 2009 cuando la prueba sufrió cambios en su estructura acorde con la Ley 1324 (Congreso de la República de Colombia, 2009) y con el Decreto 3963 (MEN, 2009a) del mismo año. Dicha reglamentación hizo que se convirtiera en obligatoria la presentación del examen para evaluar la educación formal recibida por quienes terminaban los programas de pregrado en las instituciones de educación superior (IES) y que los exámenes específicos que contaban con menos de mil evaluados fueran descontinuados, para así poder evaluar sobre competencias genéricas: “Pensamiento crítico”, “Solución de problemas”, “Entendimiento interpersonal” y “Escritura”. Condición esta que constituye la razón principal por la cual el presente libro se concentra en el periodo 2000-2009, previo a esta crucial modificación, dado que posteriormente los análisis de este tipo difícilmente fueron cotejables con dicho periodo (véanse, por ejemplo, los estudios de Pérez-Pulido et al., 2016; Melo-Becerra et al., 2017).

Posteriormente, en 2010, el Icfes inició el diseño y aplicación de un nuevo examen conocido como Saber PRO (Icfes, 2010a), el cual actualmente incluye 1) la evaluación de competencias genéricas (MEN, 2017), entendidas estas como aquellas que todos los

estudiantes deben desarrollar independiente del énfasis de formación, y 2) la evaluación de competencias comunes a grupos de programas con características de formación similares, los cuales reciben el nombre de grupos de referencia. De esta manera, se evalúan competencias genéricas en cinco pruebas: “Lectura crítica”, “Razonamiento cuantitativo”, “Comunicación escrita”, “Inglés” y “Competencias ciudadanas”, las cuales sería importante analizar y en lo posible cotejar con otros periodos, en futuros trabajos.

2.5. VARIABLES DE CONTEXTO

En el caso del presente análisis, las variables de contexto están determinadas por la asociación entre las dos pruebas censales. De esta forma, se tiene en cuenta el desempeño académico en los componentes del núcleo común de la prueba de Estado Saber 11, y su reflejo (el mismo estudiante) en un periodo después, en los componentes Saber PRO, en los cuales existe una relación biunívoca. En este caso: “Lenguaje” del Saber 11 será un determinante del componente de “Comprensión lectora” del Saber PRO y el componente “Matemáticas”, para un estudiante de Economía, se verá reflejado en el componente de “Estadística y econometría” del Saber PRO (véase concordancia con Rodríguez-Morales, 2017).

Finalmente, se debe decir que el desarrollo y efectivización de esta metodología se puede revisar en el siguiente capítulo.